様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １４ 日現在

機関番号：１７１０４
研究種目：基盤研究(B)（一般）
研究期間：2017～2021
課題番号：１７Ｈ０１７９１

研究課題名（和文）ストリームデータを知識化する圧縮情報処理基盤の開発

研究課題名（英文）Development of a compressed information processing infrastructure for converting stream data into knowledge

研究代表者
坂本　比呂志（Sakamoto, Hiroshi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：５０３１５１２３

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：大きすぎて処理できないデータは存在しないものと同義である．本研究は，データ圧縮によって情報処理を加速し，巨大なデータの理解を可能にする計算基盤を提案したい．現代は，多様で豊富なデータ，革新的なアルゴリズム，高性能なハードウェアのすべてが利用可能である．しかし，ストリームデータの激増によってこの均衡が崩れつつあり，アルゴリズムやハードウェアの飛躍的な性能向上が必要である．

研究成果の学術的意義や社会的意義
本研究は，データ圧縮を基盤とする新しい情報処理を創出しようとしている．
これまでの圧縮とその周辺技術は大まかに，(1) サイズを小さくする，(2) 圧縮データを高速照合する，(3)圧縮データを索引化する，(4) 圧縮データから類似性や頻出パターンを求めるという順に進歩してきた．しかしこれらの技術は静的データに対するものであるため現代のストリームデータには適用できない．本研究が目指す圧縮情報処理はこの弱点を克服し，既存の計算機やネットワークの処理能力をアルゴリズムによってスケールアップすることができる．

研究成果の概要（英文）：Data that is too large to process is synonymous with nonexistence. This research aims to propose a computational infrastructure that accelerates information processing by data compression and makes it possible to understand huge data. We would like to propose a computational infrastructure that accelerates information processing and enables the understanding of huge data through data compression. Today, we are faced with Today, diverse and abundant data, innovative algorithms, and high-performance hardware are all available.

研究分野：データ圧縮

キーワード：簡潔データ構造　圧縮情報検索　ストリームアルゴリズム　機械発見　秘匿計算

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
圧縮データに対して直接アクセスするための技術が 2000 年代にかけて発明され、簡潔データ構造などの圧縮情報処理が盛んに研究されるようになった。しかし、それまでの技術は、静的なデータを対象とするものがほとんどであり、ダイナミックに変化するデータに対する圧縮情報処理の研究はほとんど行われてこなかった。しかし、ネットワーク上を流れるストリームデータ上で圧縮情報処理を実現するには、これまでにない新しい理論とその活用が必要である。また、データの多様性が高まるにつれて、個人情報を含んだデータがますます増大しており、プライバシー保護を保証しながら様々な計算を可能にする秘匿計算の需要も高まっている。

２．研究の目的
主として、テキストデータなどの非定型データが与えられた時、そのデータに対する基本演算を可能にする簡潔データ構造を構築することができる。この機構によって、データを圧縮した状態で、復号することなく、データに対する情報検索や頻出パターン発見を可能にする。しかし、従来の手法では、データ全体をいったんメモリに格納してから圧縮構造に変換するため、圧縮前後のデータサイズが極端に異なるビッグデータに対しては対応できない。また、ビッグデータの多くは、ネットワークを流れるストリームデータであるため、データ全体をいったん記憶することができない。このような理由からストリームデータに対応可能な簡潔データ構造のアルゴリズムとそれをさまざまな実用的な課題（パターン発見や機械学習）に適用可能にする理論の構築を目指す。また、個人情報を含むデータに対しても本研究のアルゴリズムを適用可能にするために、プライバシー保護計算との協調を目指して、開発環境を整備する。

３．研究の方法
（１）簡潔データ構造の理論のオンラインアルゴリズム
従来の簡潔データ構造は、静的なデータに対するオフラインアルゴリズムである。このアルゴリズムをデータがダイナミックに変化する場合に拡張することは容易ではない。一般に、簡潔データ構造が表現するデータに対する、任意の挿入と削除を許容すると、対応する簡潔データ構造は、全く異なるものとなるため、最悪の場合には、データ全体の更新が必要になるので現実的ではない。そこで、本研究では、データの更新が末尾に追加されることに限定することで、簡潔データ構造のための効率的なオンラインアルゴリズムを提案する。ストリームデータの性質からこのような制約は自然である。

（２）プライバシー保護計算への応用
テキストデータには個人情報を多く含むものが少なくない。これらのデータから個人名を削除することが考えられるが、それらの匿名処理は個人情報を秘匿する観点からは効果が薄いことが知られている。また、ゲノム情報のように匿名化ができないデータも存在する。そこで、これらのデータを秘匿したまま必要な情報を取り出す秘匿検索による手法が注目されている。本研究では、非定型データ上の圧縮索引を応用した、パターン発見アルゴリズムを構築し、それを秘匿計算可能にする機構へ拡張する。プライバシー保護計算は大きく分けて、秘密分散による手法と準同型暗号に基づく手法があるが、委託計算が可能となる後者の手法を採用する。

（３）機械学習への応用
圧縮データから機械学習することで、記憶領域やスピードアップが可能となることが知られている。しかし、これまでの研究は、画像データに対する学習に限られていた。そこで、本研究では、テキストデータからの機械学習においてもこのような性質が成り立つことを示す。具体的には、機械翻訳の学習モデルへ圧縮データを学習データとして取り入れることで、翻訳精度が向上することを明らかにする。

４．研究成果
（１）文法圧縮に基づく動的な簡潔データ構造の理論と周辺技術
文法圧縮はテキストを生成する適切な文脈自由文法を構築することでデータの冗長性を削減する可逆圧縮の一つである。この圧縮の枠組みに対して、オンラインアルゴリズムを構築し、理論的に最小のスペースで圧縮可能であることを示した[1]。また、これらの手法をさらに拡張して、より実用的なアルゴリズムを提案した[2,3,4,5]。
[1] Yoshimasa Takabatake and Tomohiro I and Hiroshi Sakamoto, A Space-Optimal Grammar Compression, LIPIcs 87, 67:1--67:15, 2017
[2] Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto, LZ-ABT: A Practical Algorithm for α-Balanced Grammar Compression, LNCS 10979, 323-335, 2018
[3] Kensuke Sakai, Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I,

Hiroshi Sakamoto, RePair in Compressed Space and Time, Data Compression Conference 2019
[4] Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Yoshimasa Takabatake, pair: Rescaling RePair with Rsync, LNCS11811, 35-44, 2019
[5] Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Louisa Seelbach Benkner, Yoshimasa Takabatake, Practical Random Access to SLP-Compressed Texts, LNCS 12303, 221-231, 2020

（２）BWTに基づく動的な簡潔データ構造の理論と周辺技術
簡潔データ構造と親和性の高い圧縮アルゴリズムとして、文法圧縮とは独立の技術として、BWT と呼ばれる変換アルゴリズムがある。これはテキストデータをより圧縮がしやすいデータへ並べ替えるための手法で、それ自身は圧縮ではない。テキストデータを BWT によって変換することで、きわめて圧縮が容易な形式に変換されるが、BWT は計算コストが高いことや、オンライン構築が困難であることが知られている。そこで本研究では、データの末尾追加という制約を課すことで、これらの問題点を解決した[6]。また BWT から圧縮手法への変換を効率的に実行する実用的なアルゴリズムを提案した[7]。
[6] Tatsuya Ohno, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto, A Faster Implementation of Online Run-Length Burrows-Wheeler Transform, LNCS 10765, 409-419, 2018
[7] Tatsuya Ohno, Kensuke Sakai, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto, A faster implementation of online RLBWT and its application to LZ77 parsing, J. Discrete Algorithms, 52-53, 18-28, 2018

（３）テキストデータ上の編集距離の計算とその応用
文法圧縮は、フレーズ単位で圧縮することで、構文木に変換する。したがって、テキストデータに繰り返し構造が多く出現するほど、圧縮率が高くなる。この性質を利用して、データ中の繰り返し構造の発見や、構文木の類似度を比較することで、頻出パターン発見に応用できることを示した[8]。このデータベースを arxiv などのプレプリントサーバ上の論文データに対して適用することで、非常に類似した剽窃が疑われる論文を発見できることを実証した。
[8] Shouhei FUKUNAGA, Yoshimasa TAKABATAKE, Tomohiro I, Hiroshi SAKAMOTO, Approximate Frequent Pattern Discovery in Compressed Space, IEICE Transactions on Information and Systems E101.D (3), 593-601, 2018

（４）圧縮データによる機械学習の性能向上
機械翻訳の精度を飛躍的に高める前処理として、サブワード分解と呼ばれる手法が知られている。これは、通常の言語による品詞分解ではなく、機械学習にとって都合のよい分解を行うことで、学習時の未知語の増加を抑える効果が期待できるためである。これまでに様々なサブワード分解が提案されているが、本研究では、最新の文法圧縮によるサブワード分解を提案し、特に訓練データが少ない場合にこれまでの手法を上回る精度で学習が可能となることを示した[9]。また、Jpeg データの符号化で行われている DCT 変換係数から直接学習が可能であることを示し、この手法を応用して画像生成を可能にした[10,11]。
[9] Keita Nonaka, Kazutaka Yamanouchi, Tomohiro I, Tsuyoshi Okita, Kazutaka Shimada, Hiroshi Sakamoto, A Compression-Based Multiple Subword Segmentation for Neural Machine Translation, Electronics, 11(7), 1014, 2022
[10] 大北 剛，管谷克彦，坂本 比呂志，JPEG の画像表現を用いた画像生成の高速化，第 24 回情報論的学習理論ワークショップ(IBIS2021)，2021
[11] 管谷 克彦，高畠 嘉将，井 智弘，申 吉浩，坂本 比呂志，非可逆圧縮データからの高速な画像生成，第 23 回情報論的学習理論ワークショップ(IBIS2020)，2020

（５）プライバシー保護計算への応用
プライバシー保護計算の枠組みとして、本研究では、準同型暗号に基づく手法を構築する。準同型暗号は、暗号化した整数に対する加算や乗算を可能にする枠組みであり、これらを組み合わせることで様々な計算を可能にする。しかし、計算コストが高いため、どのような問題に適用するか、その時のコストをどのように削減するかが重要である。本研究では、これまでに構築したデータ同士の編集距離を計算する 2 パーティモデルを提案した[12,13]。またこれらの手法に対する知財を獲得した[14]。さらに今後の開発を容易にするために、準同型暗号のための多機能ライブラリを C++上で実装して、GitHub 上に公開している[15]。これまでは、加算や乗算などの論理回路を最初から構築する必要があったが、このライブラリによって準同型暗号上の基本演算をこれらの関数を呼び出すことで簡単に利用できるようになった。
[12] Shunta Nakagawa, Tokio Sakamoto, Yoshimasa Takabatake, Tomohiro I, Kilho Shin, Hiroshi Sakamoto, Privacy-Preserving String Edit Distance with Moves, LNCS 11223, 226-

240, 2018
[13] Yohei Yoshimoto, Masaharu Kataoka, Yoshimasa Takabatake, Tomohiro I, Kilho Shin, Hiroshi Sakamoto, Faster Privacy-Preserving Computation of Edit Distance with Moves, LNCS 12049, 308-320, 2020
[14] 秘匿検索システム及び秘匿検索プログラム，坂本比呂志　申吉浩，特願2019-93908，2019
[15] 拡張 TFHE ライブラリ（坪田優希，坂本比呂志）
https://github.com/hiroshi-kyutech/FTHE-tool

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Louisa Seelbach Benkner, Yoshimasa Takabatake	４．巻 12303
２．論文標題 Practical Random Access to SLP-Compressed Texts	５．発行年 2020年
３．雑誌名 LNCS	６．最初と最後の頁 221-231
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-59212-7_16	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Yohei Yoshimoto, Masaharu Kataoka, Yoshimasa Takabatake, Tomohiro I, Kilho Shin, Hiroshi Sakamoto	４．巻 12049
２．論文標題 Faster Privacy-Preserving Computation of Edit Distance with Moves	５．発行年 2020年
３．雑誌名 LNCS	６．最初と最後の頁 308-320
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-39881-1_26	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yohei Yoshimoto, Masaharu Kataoka, Yoshimasa Takabatake, Tomohiro I, Kilho Shin, Hiroshi Sakamoto	４．巻 LNCS12049
２．論文標題 Faster Privacy-Preserving Computation of Edit Distance with Moves	５．発行年 2020年
３．雑誌名 WALCOM2020	６．最初と最後の頁 308-320
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-39881-1_26	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Yoshimasa Takabatake	４．巻 LNCS11811
２．論文標題 pair: Rescaling RePair with Rsync	５．発行年 2019年
３．雑誌名 SPIRE 2019	６．最初と最後の頁 35-44
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-32686-9_3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Tatsuya Ohno, Kensuke Sakai, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto	4．巻 52-53
2．論文標題 A faster implementation of online RLBWT and its application to LZ77 parsing	5．発行年 2018年
3．雑誌名 J. Discrete Algorithms	6．最初と最後の頁 18-28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jda.2018.11.002	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto	4．巻 10979
2．論文標題 LZ-ABT: A Practical Algorithm for α-Balanced Grammar Compression	5．発行年 2018年
3．雑誌名 Lecture Notes in Computer Science	6．最初と最後の頁 323-335
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Shunta Nakagawa, Tokio Sakamoto, Yoshimasa Takabatake, Tomohiro I, Kilho Shin, Hiroshi Sakamoto	4．巻 11223
2．論文標題 Privacy-Preserving String Edit Distance with Moves	5．発行年 2018年
3．雑誌名 Lecture Notes in Computer Science	6．最初と最後の頁 226-240
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Shouhei FUKUNAGA, Yoshimasa TAKABATAKE, Tomohiro I, Hiroshi SAKAMOTO	4．巻 E101.D (3)
2．論文標題 Approximate Frequent Pattern Discovery in Compressed Space	5．発行年 2018年
3．雑誌名 IEICE Transactions on Information and Systems	6．最初と最後の頁 593-601
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2017FCP0010	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Yoshimasa Takabatake and Tomohiro I and Hiroshi Sakamoto	4．巻 87
2．論文標題 A Space-Optimal Grammar Compression	5．発行年 2017年
3．雑誌名 LIPIcs	6．最初と最後の頁 67:1--67:15
掲載論文のDOI（デジタルオブジェクト識別子） 10.4230/LIPIcs.ESA.2017.67	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Tatsuya Ohno, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto	4．巻 10765
2．論文標題 A Faster Implementation of Online Run-Length Burrows-Wheeler Transform	5．発行年 2018年
3．雑誌名 LNCS	6．最初と最後の頁 409-419
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-78825-8_33	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　3件／うち国際学会　2件）

1．発表者名 管谷 克彦，高畠 嘉将，井 智弘，申 吉浩，坂本 比呂志
2．発表標題 非可逆圧縮データからの高速な画像生成
3．学会等名 IBIS2020
4．発表年 2020年

1．発表者名 Kensuke Sakai, Tatsuya Ohno, Keisuke Goto, Yoshimasa Takabatake, Tomohiro I, Hiroshi Sakamoto
2．発表標題 RePair in Compressed Space and Time
3．学会等名 Data Compression Conference 2019（国際学会）
4．発表年 2019年

１．発表者名
Hiroshi Sakamoto

２．発表標題
Stream Data Compression and Its Applications

３．学会等名
The 14th International Conference on Modeling Decisions for Artificial Intelligence（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
坂本比呂志

２．発表標題
ストリームデータ圧縮とその応用

３．学会等名
第30回 回路とシステムワークショップ プログラム（招待講演）

４．発表年
2017年

１．発表者名
坂本比呂志

２．発表標題
データ圧縮の機械学習と秘匿計算への応用

３．学会等名
第120回人工知能基本問題研究会（招待講演）

４．発表年
2022年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称	発明者	権利者
秘匿検索システム及び秘匿検索プログラム	坂本比呂志　申吉浩	同左
産業財産権の種類、番号	出願年	国内・外国の別
特許、特願2019-93908	2019年	国内

〔取得〕　計0件

〔その他〕

拡張TFHEライブラリ
https://github.com/hiroshi-kyutech/FTHE-tool
完全準同型暗号(TFHE)を拡張した高機能C++ライブラリを開発し、Gighub上で公開した。

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	竹田　正幸 (Takeda Masayuki) (50216909)	九州大学・システム情報科学研究院・教授 (17102)	
研究分担者	申　吉浩 (Shin Yoshihiro) (60523587)	学習院大学・付置研究所・教授 (32606)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関